Using linguistic typology to enrich multilingual lexicons: the case of lexical gaps in kinship

NATIONAL UNIVERSITY OF MONGOLIA

Temuulen Khishigsuren, Gábor Bella, Khuyagbaatar Batsuren, Abed Alhakim Freihat, Nandu Chandran Nair, Amarsanaa Ganbold, Hadi Khalilia, Yamini Chandrashekar, Fausto Giunchiglia

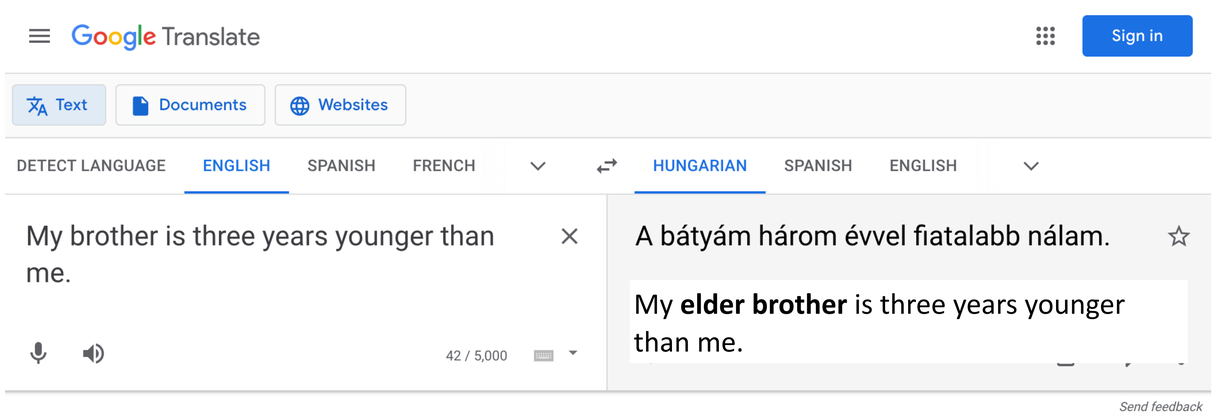
UNIVERSITY OF TRENTO

Why is lexical diversity important?

Use of **typology i**n various NLP applications generated consistent improvements; hence, regarded as a promising new direction to tackle the issue of data scarcity in multilingual NLP [1].

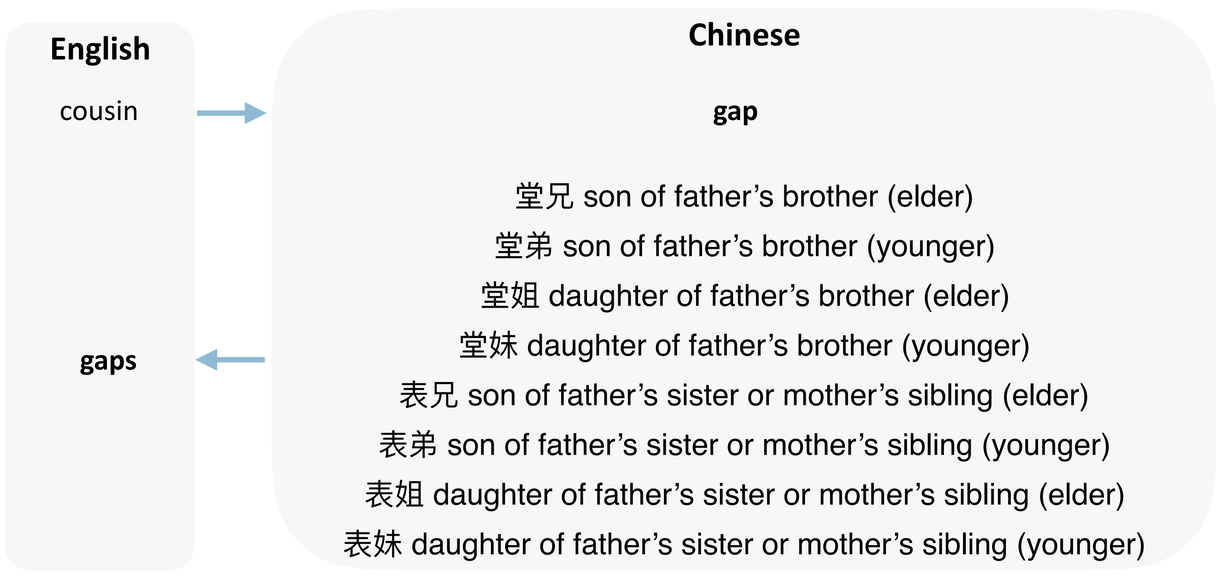
However, most typology-informed NLP studies are limited to morphosyntactic features and have so far ignored **lexical diversity.**

Ignoring diversity in lexicons can lead to hard-to-detect meaning-level mistakes:



What are lexical gaps?

A concept is considered a lexical gap if it can only be expressed through free combinations of words [2].



Methods: How to use lexical typology to infer lexical gaps?

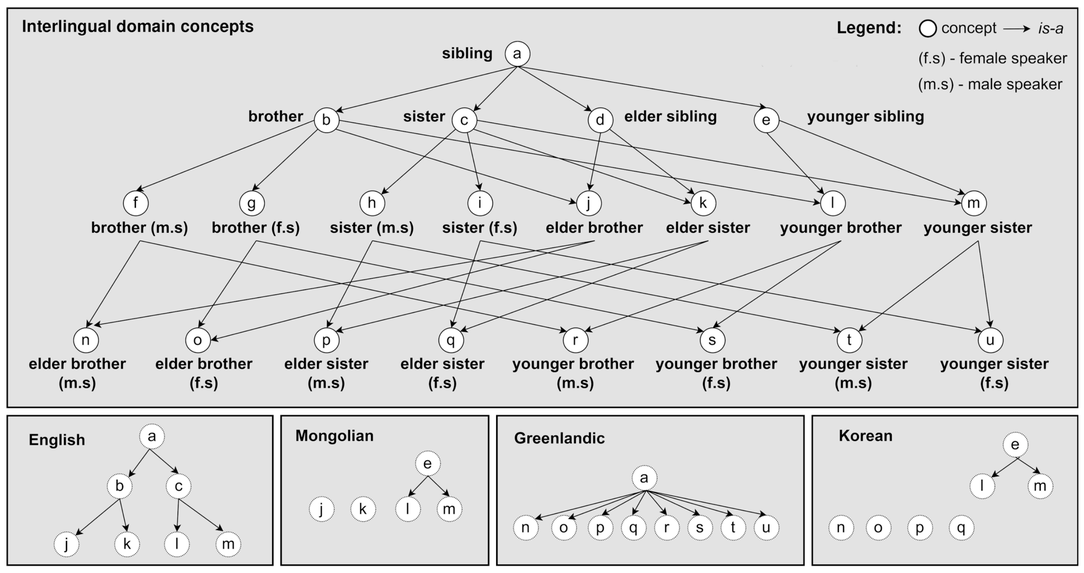
(1) Data collection:

Typological knowledge:
- Murdock's lexicalization patterns [3]
- 566 languages
- 4 subdomains of kinship

Existing resources:
- Wiktionary
- 166 languages, 1681 words
- 6 subdomains of kinship

Native speakers:
- 10 languages, 230 words
- 6 subdomains of kinship

(2) Conceptual modeling:



(3) Gap inference:

For concepts with speaker gender and age undefined:

if neither a concept c nor its parents have a lexicalization in language ℓ then c is a lexical gap in ℓ.

For concepts with speaker gender or age specified:

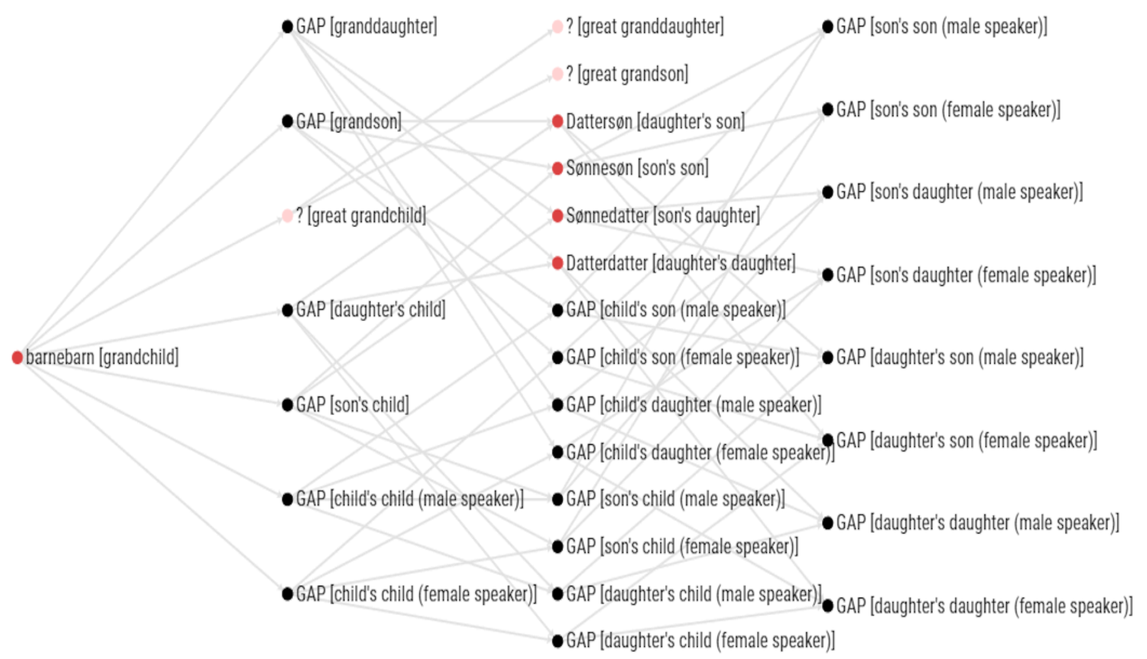
language ℓ is known not to indicate the speaker's gender or age in the lexicalization, then all concepts with these attributes are lexical gaps in ℓ.

Published resource: Availability and visibility

Freely accessible as a stand-alone resource at: http://github.com/kbatsuren/KinDiv
http://ukc.disi.unitn.it/index.php/kinship/

Can be browsed and visualized at: http://ukc.datascientia.eu
http://www.livelanguage.eu

Domain	Concepts	Languages	Words	Gaps
grandparents	19	539	391	7,171
grandchildren	27	247	202	5,049
siblings	21	304	498	3,851
uncles&aunts	31	625	312	16,503
nephews&nieces	33	65	214	1,606
cousins	67	60	294	3,190
Total	**198**	**699**	**1,911**	**37,370**



Contact

Email: **kh.temulen@gmail.com, khuyagbaatar@num.edu.mn**
Address: National University of Mongolia
Ikh surguuliin gudamj-1, Sukhbaatar district
Ulaanbaatar, Mongolia

References

[1] Ponti, E.M., O'horan, H., Berzak, Y., Vulić, I., Reichart, R., Poibeau, T., Shutova, E. and Korhonen, A., 2019. Modeling language variation and universals: A survey on typological linguistics for natural language processing. *Computational Linguistics*, *45*(3), pp.559-601.

[2] Pianta, E., Bentivogli, L. and Girardi, C., 2002. MultiWordNet: developing an aligned multilingual database. In *First international conference on global WordNet* (pp. 293-302).

[3] Murdock, G.P., 1970. Kin term patterns and their distribution. *Ethnology*, *9*(2), pp.165-208.